DOCUMENT RESUME

ED 325 492 TM 015 709

AUTHOR Kaiser, Javaid
TITLE The Robustness of Regression and Substitution by Mean Methods in Handling Missing Values.
PUB DATE Aug 90
NOTE 24p.; Paper presented at the Annual Islamic Conference on Statistical Sciences (2nd, Johor Bahru, Malaysia, August 26-30, 1990).
PUB TYPE Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; Computer Simulation; Estimation (Mathematics); *Predictor Variables; *Regression (Statistics); Research Problems; *Robustness (Statistics); Sample Size; Surveys
IDENTIFIERS Covariance Structural Analysis; *Missing Data; Survey Research

ABSTRACT

There are times in survey research when missing values need to be estimated. The robustness of four variations of regression and substitution by mean methods was examined using a 3x3x4 factorial design. The regression variations included in the study were: (1) regression using a single best predictor; (2) two best predictors; (3) all available preditors having observed values; and (4) all available predictors with adjustment of estimate for predictors having missing values. The factors studied included sample size (n=30, 60, and 120), the proportion of incomplete records (IRs) in the sample (IR=10%, 20%, and 30%), and the number of missing values (MVs) per IR (MVs=12.5%, 25%, 37.5%, and 50%). The design matrix was replicated 500 times. Imputation methods were compared in terms of retaining population covariance structure in imputed samples. The results suggest that all methods significantly altered covariance structure and that the regression variation that adjusts missing value estimates for predictors having missing values was found to be the best imputation method at all experimental conditions. Three data tables and six graphs are included. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED325492

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JAVAID KAISER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

THE ROBUSTNESS OF REGRESSION AND SUBSTITUTION BY MEAN METHODS IN HANDLING MISSING VALUES

Javaid Kaiser
Virginia Polytechnic Institute & State University

Presented at the Second Islamic Countries Conference on Statistical Sciences, Johor Bahru, Malaysia
August 26-30, 1990

TM015709

THE ROBUSTNESS OF REGRESSION AND SUBSTITUTION BY MEAN METHODS IN HANDLING MISSING VALUES

JAVAID KAISER
College of Education-AES, Virginia Tech.,
Blacksburg, VA 24061-0302, USA

Abstract

The robustness of four variations of regression and substitution by mean method was studied using 3x3x4 factorial design. Imputation methods were compared in terms of retaining population covariance structure in imputed samples. The results suggest that all methods significantly altered covariance structure and that the regression variation that adjusts missing value estimates for predictors having missing values was found the best imputation method at all experimental conditions.

KEYWORDS

Regression Missing Values Simulation Zero-order

There are times in survey research when missing values need to be estimated. The regression and substitution by mean (zero order) methods are the most commonly used techniques, for this purpose. Regression method was first proposed by Buck (1960). In this method, the variable having missing value is treated as a criterion variable and is regressed on variables having observed values to predict the criterion variable. There are several variations to the regression method depending on how many predictors are used in prediction and how missing values on the predictors themselves, are handled (Kim & Kohout, 1975). The selection of predictors depends on the correlation of predictors with the criterion variable. Theoretically, a large set of predictors should produce a better estimate of the missing value. However, use of too many predictors overfits the regression equation and causes poor estimates (Frane, 1976). Computation of initial correlation matrix from an incomplete data matrix is also controversial. (Buck, 1960; Gleason & Staelin, 1975; Timm, 1970).

The zero-order method was first introduced by Wilks (1932) and is widely criticized for distorting the distribution by creating spikes and reducing variance estimates (Kalton & Kish, 1981; Proctor, 1978). It is generally recommended when variables in the data matrix are not highly correlated with the variable having missing value (Afifi & Elashoff, 1967).

The purpose of estimating missing values is very important to consider while selecting an imputation method. If the intent is simply to estimate population means on variables represented in the data matrix, a method that is less expensive and involves less computations can easily be found. If the purpose is to estimate missing values to complete individual records, the quality of every single estimate of missing value is important. In this event, procedure used by Geason et al. (1975) may be used to determine the quality of missing value estimates. When missing values are imputed to complete the data matrix with the intention to test hypothesis, one should pay close attention to the covariance structure of the matrix. Missing value estimates tend to alter the covariance structure of the data. This is a serious drawback in imputing missing values and needs careful examination.

No imputation method, particularly the ones discussed above, are robust for all purposes of estimation but are used indiscriminately. At present, the selection of imputation method is primarily a matter of taste and depends on the statistical background of the researcher. The need for an empirical investigation to determine the robustness of regression and zero-order methods was strongly felt. This study is a step forward in this direction.

Method

The robustness of four variations of regression method and zero order method was investigated in terms of their retaining population covariance structure in imputed samples. The regression variations included in the study were regression using: (1) single best predictor, (2) two best predictors, (3) all available predictors having observed values, and (4) all available predictors with adjustment of estimate for predictors having missing values. The study was conducted with 3x3x4 factorial design. The factors studied were sample size (n = 30, 60, 120), the proportion of incomplete records in the sample (IR = 10%, 20%, 30%), and the number of missing values per incomplete record (MV = 12.5%, 25%, 37.5%, 50%). The design matrix was replicated 500 times.

Data matrices of multivariate normal deviates of size nx9 were generated at random from a known population covariance matrix, given in Table 1. The first variable was named exogenous variable (V1) and was exclusively used to create systematically missing values. The remaining nx8 matrix was used for imputation. The covariance structure of every matrix generated was tested against the known population covariance using the equation

$$-2\log\lambda = pn\,(\log n - 1) - n \log |B\psi^{-1}| + \mathrm{tr}\,(B\psi^{-1})$$

where p = number of variables in the matrix
n = sample size
B = sum of squares and sum of products matrix
ψ = population variance-covariance matrix

as suggested by Anderson (1958). The test statistic $-2\log\lambda$ is asymptotically distributed as chi-square distribution with $p(p+1)/2$ degrees of freedom. The data matrices that had covariance structure similar to that of the population ($p > .05$) were retained for use in this study. Five hundred matrices were generated for each cell of the design matrix.

Missing values occurring systematically were artificially created using the following model. A high value on the exogenous variable (V1) caused the first and third variables to show missing values. The variables 5 and 8 showed missing values whenever observed value on variable 3 exceeded .4. Variables 6 and 7 showed missing values when the observed value on variable 3 was .4 or less. Once the missing values having a systematic pattern of occurrence were created, the imputation methods were used one at a time to impute missing values. The imputed matrices were tested against population covariance structure by using equation described above at .10, and .05 levels of significance. The statistic showing number of matrices that could not retain population covariance structure because of imputation was compiled for all the methods over all replications.

Results

Table 2 presents the results in terms of the percent of data matrices that could not retain population covariance structure at .05 level of significance. Figures 1, 2, and 3 present the same information in graphic form for sample sizes of 30, 60, and 120. The data indicated that all imputation methods altered the covariance structure of the imputed samples to some degree and that data matrices with significantly altered covariance increased as the number of incomplete records in a sample or the number of missing values per incomplete record increased. The number of such matrices ranged from 4.2 to more than 99 percent.

The study also revealed that the regression variations representing one predictor (REGONE), two predictors (REGTWO), and all predictors (REGALL) caused more damage to the covariance structure than zero-order and REGRESS methods. This finding was true for all levels of sample sizes, proportion of incomplete records (IR), and the number of missing values (MV). However the differences within REGONE, REGTWO, and REGALL diminished as the sample size increased. At sample size of 120, there seemed no difference in the three techniques. In samples of size 30, zero-

order and REGRESS were equally efficient in retaining population covariance structure in imputed samples at all levels of IR and MV. When 30% of the records were incomplete, the efficiency of the two methods varied at various levels of missing values but no systematic trend was found. There was no difference between zero-order and REGRESS for all levels of missing values when the proportion of incomplete records was 10%. However, as the sample size increased, REGRESS performed better than the zero-order method. Overall, the REGRESS was considered the best imputation technique that retained population covariance structure after imputing missing values.

Data matrices that did not retain population covariance structure at .10 level of significance is given in Table 3. The same information is presented in Figures 4, 5, and 6 for sample size of 30, 60, and 120. All the findings obtained at .05 level of significance were also found true at .10 level except that more matrices having significantly different covariance than that of the population were identified.

The results of this study suggest that one should pay attention to the covariance structure of the data matrix after imputing missing values particularly when values are missing systematically. No imputation procedure amongst the ones studied here was found satisfactory in terms of retaining population covariance structure. All methods produced more matrices with significantly altered covariance than suggested by the alpha level. However, it was very clear that for small sample sizes like 30 or when the proportion of incomplete records is not more than 10%, zero-order and REGRESS methods are equally good candidates and that the former may be preferred because of low computing cost. In larger samples or when IR exceeds 10%, the REGRESS was distinctly a method of choice. The implications of these findings are very serious because the researcher may end up completing the data matrix at the cost of losing representativeness of population covariance which in turn may question the conclusions drawn from such imputed matrix.

Conclusion

The results revealed that REGRESS is the best method of imputation in terms of retaining population covariance structure when missing values occur systematically. Substitution by mean method was considered an equally a good candidate for samples of size 30 when the proportion of incomplete records is 10% or below. One predictor, two predictors, and all predictors variations of regression were found less efficient than REGRESS at all levels of all treatment conditions. Besides, these three variations were not found different from one another in terms of retaining population covariance structure. As the sample size increased, these differences became less and less significant.

In terms of general trends, the number of matrices having significantly altered covariance structure increased with the increase in the proportion of incomplete records in the sample or when the number of missing values increased within a given level of IR.

REFERENCES

Afifi, A. A., & Elashoff, R. M. (1967). Missing observations in multivariate statistics II. Point estimation in simple linear regression. Journal of American Statistical Association, 62 (317) 10-29.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. Journal of the Royal Statistical Society, Series B, 22, 302-307.

Frane, J. W. (1976). Some simple procedures for handling missing data in multivariate analysis. Psychometrika, 41, (3), 409-415.

Gleason, T. C., & Staelin, R. (1975). A proposal for handling missing data. Psychometrika, 40(2), 229-252.

Kalton, G., & Kish, L. (1981). Two efficient random imputation procedures. Proceedings of the Survey Research Section, American Statistical Association, 146-151.

Kim, J. O., & Kohout, F. J. (1975). Multiple regression analysis: Subprogram regression. In N. H. Nie, C. Hadlaihull, J. G. Jenkins, K. Steinbrenner, & D. H. Bent (Eds.), Statistical Package for Social Sciences (2nd. Ed.). New York: McGraw-Hill Book Co., 347-348.

Proctor, C. H. (1978). More on imputing versus deleting when estimating scale scores. Proceedings of the Survey Research Section, American Statistical Association.

Timm, N. H. (1970). The estimation of variance-covariance and correlation matrices from incomplete data. Psychometrika, 35 (4), 417-438.

Wilks, S. S. (1932). Moments and distributions of population parameters from fragmentary samples. Annals of Mathematical Statistics, 3, 163-195.

Table 1

Population Correlation Matrix

	V1	V2	V3	V4	V5	V6	V7	V8
V1	1.00							
V2	.318	1.00						
V3	.468	.230	1.00					
V4	.403	.317	.305	1.00				
V5	.321	.285	.247	.227	1.00			
V6	.414	.272	.263	.322	.187	1.00		
V7	.365	.292	.297	.339	.398	.388	1.00	
V8	.413	.232	.250	.380	.441	.283	.463	1.00

Table 2

The Percent of Data Matrices that Could not Retain Population Covariance Structure at .05 Level

	IR	10%				20%				30%			
	MV	1	2	3	4	1	2	3	4	1	2	3	4
n=30	Zero-order	6.60	8.20	9.40	10.20	12.60	16.20	23.00	29.00	20.40	25.80	44.40	54.80
	One predictor	7.40	12.00	16.40	19.60	17.20	28.00	38.20	47.80	27.40	45.20	67.40	81.60
	Two predictors	8.60	13.60	19.60	25.20	18.20	32.00	45.40	58.80	31.80	58.20	78.80	91.20
	All predictors	12.00	18.00	23.80	27.80	27.20	41.20	57.00	64.60	38.60	72.60	87.40	79.00
	REGRESS	6.00	8.00	10.00	12.40	10.00	15.40	21.40	33.20	15.20	26.60	34.60	57.40
n=60	Zero-order	7.00	6.40	8.40	11.20	15.60	20.40	30.80	38.80	31.40	44.00	62.60	76.40
	One predictor	6.80	10.40	15.00	18.00	15.60	28.60	46.40	56.80	34.80	60.60	82.40	90.60
	Two predictors	6.60	12.00	16.60	20.80	16.40	34.00	50.80	63.80	35.40	67.20	87.00	94.20
	All predictors	9.00	14.40	19.00	23.20	21.60	42.00	55.60	66.20	41.60	76.20	89.00	96.00
	REGRESS	4.40	5.00	7.20	10.80	8.00	12.60	24.20	34.00	16.80	28.80	51.00	73.40
n=120	Zero-order	9.20	11.20	13.80	18.40	29.00	38.40	53.00	61.80	59.20	76.00	91.00	97.40
	One predictor	9.40	13.40	19.20	23.40	28.00	45.40	61.80	76.00	56.40	83.60	93.20	98.40
	Two predictors	9.40	13.80	21.00	25.20	26.00	47.20	63.40	77.60	54.60	85.40	96.60	99.20
	All predictors	8.40	14.00	21.00	27.00	24.20	48.00	65.60	80.20	53.00	86.00	96.40	99.00
	REGRESS	6.00	7.80	10.60	14.00	11.80	22.00	34.20	48.60	27.80	55.00	79.00	91.00

IR: Percent of incomplete records per sample
MV: Number of missing values per IR

Table 3

The Percent of Data Matrices that Could not Retain Population Covariance Structure at .10 Level

	IR	10%				20%				30%			
	MV	1	2	3	4	1	2	3	4	1	2	3	4
n=30	Zero-order	17.20	17.60	19.80	20.60	23.00	27.20	35.80	38.20	30.20	37.20	56.20	68.80
	One predictor	19.20	23.00	27.60	33.40	29.20	37.20	51.60	62.80	39.20	60.20	79.40	91.00
	Two predictors	19.40	25.80	31.80	38.00	31.80	43.20	60.20	72.80	45.00	69.60	87.40	95.20
	All predictors	22.80	32.00	37.60	39.80	39.40	56.00	68.60	76.60	47.40	82.40	93.60	86.00
	REGRESS	15.20	17.00	18.60	22.80	19.60	27.20	30.60	44.40	23.20	37.80	48.00	69.40
n=60	Zero-order	14.00	14.40	18.60	20.00	26.80	32.40	45.20	52.00	40.40	57.00	74.80	86.80
	One predictor	15.60	20.00	25.40	29.60	31.20	43.60	57.00	71.80	48.40	74.80	88.40	96.60
	Two predictors	15.80	20.60	28.20	33.40	31.00	47.20	65.60	76.00	47.80	79.40	91.40	97.40
	All predictors	17.00	24.00	31.80	36.00	34.40	53.80	69.60	78.60	56.00	84.40	93.80	97.60
	REGRESS	11.00	12.00	17.00	19.80	18.60	22.40	35.00	47.80	27.20	44.80	66.80	82.00
n=120	Zero-order	19.40	20.60	25.60	27.40	39.80	51.40	64.80	73.00	70.20	85.80	95.40	98.60
	One predictor	18.00	26.20	31.80	35.40	40.60	59.00	75.20	83.80	70.00	92.20	98.20	99.40
	Two predictors	17.20	25.60	32.80	36.80	38.40	60.80	78.00	86.00	68.20	92.20	98.00	99.60
	All predictors	17.20	28.40	35.40	38.40	37.80	62.40	80.00	87.20	66.40	91.60	98.40	99.80
	REGRESS	12.40	14.20	20.40	24.60	22.00	34.40	48.00	62.80	40.80	69.20	87.00	95.20

IR: Percent of incomplete records per sample
MV: Number of missing values per IR

Figure 1

The Percent of Data Matrices that did not Retain Population Covariance Structure for N=30 at .05 Level of Significance

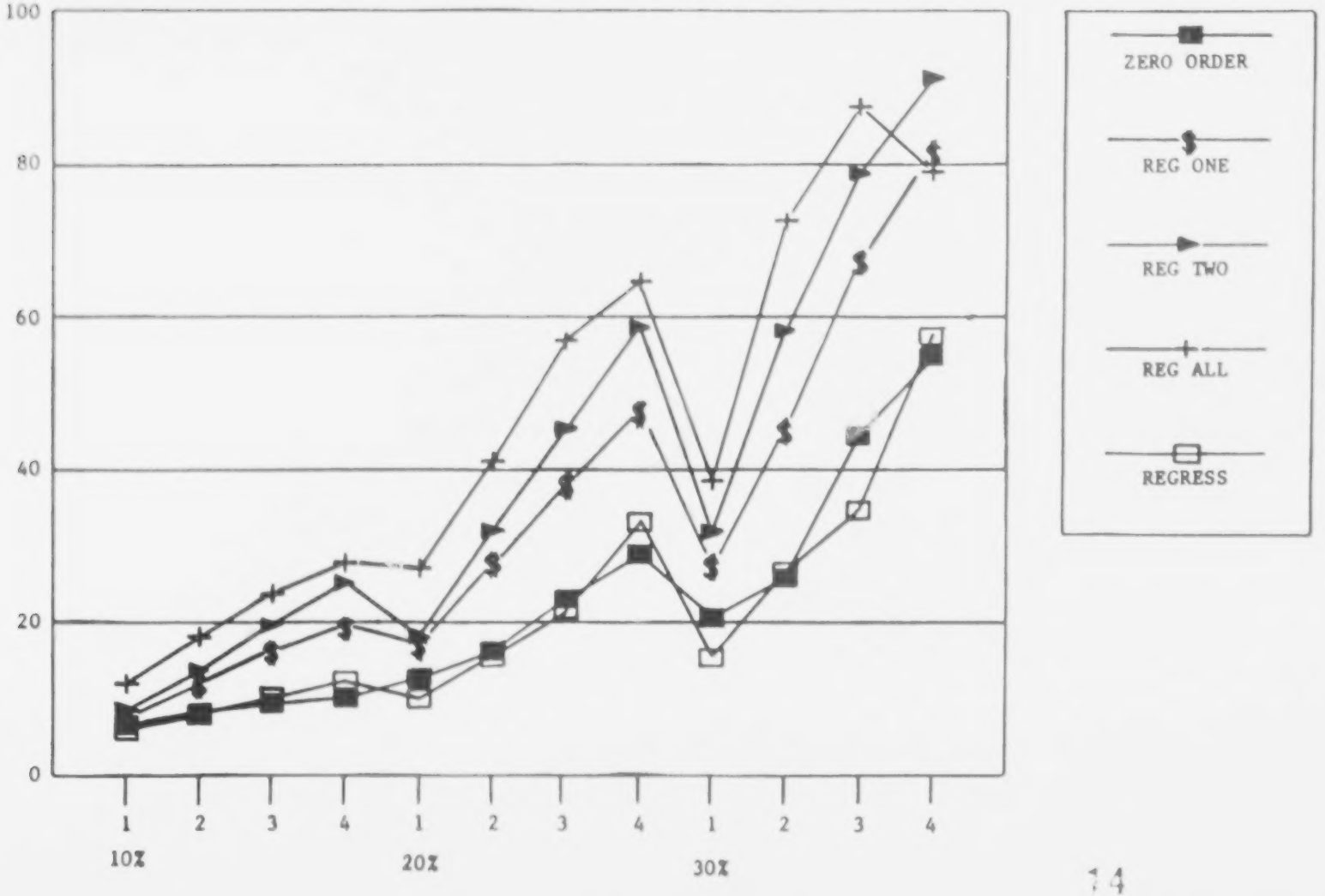

Figure 2

The Percent of Data Matrices that did not Retain Population Covariance Structure for N=60 at .05 Level of Significance

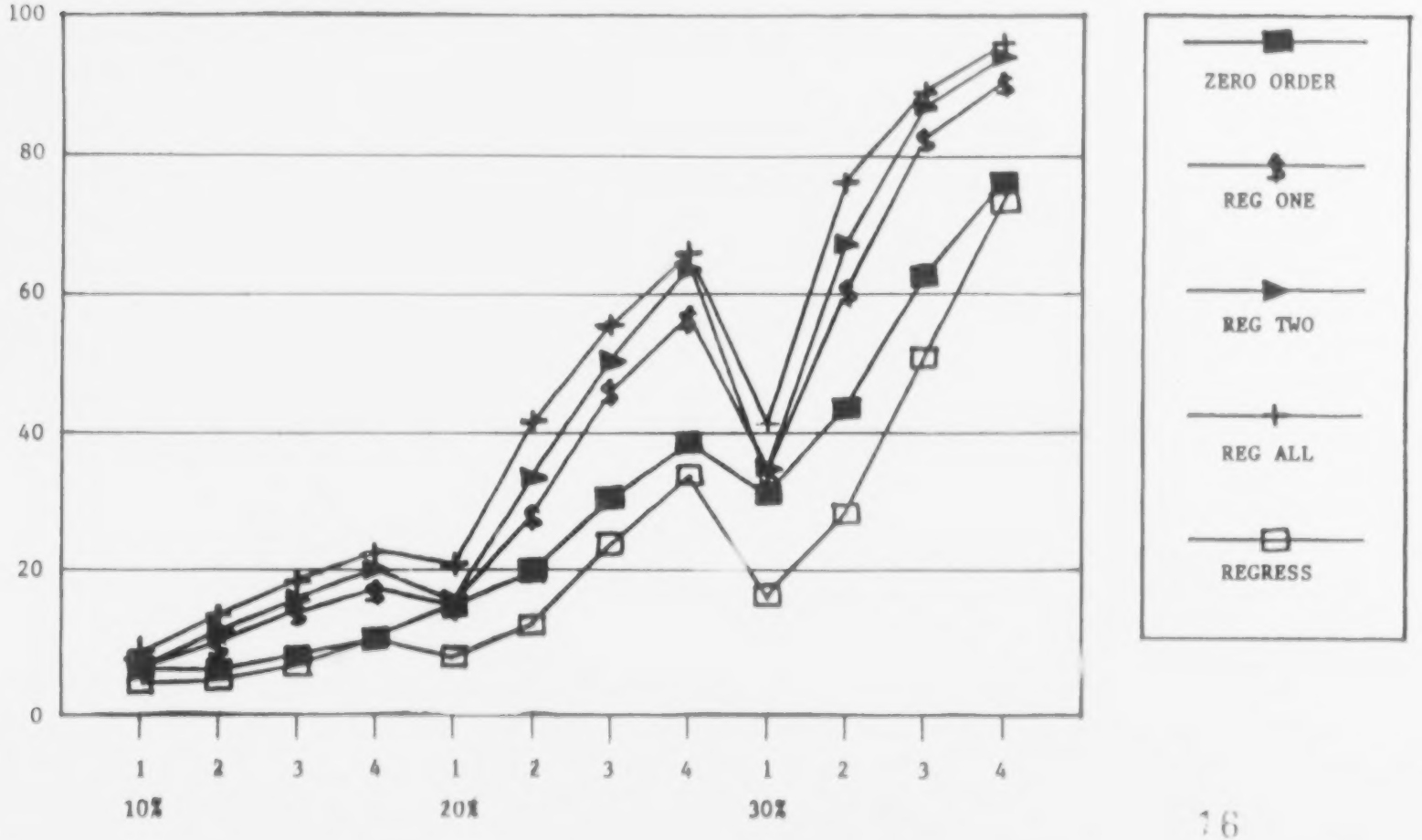

Figure 3

The Percent of Data Matrices that did not Retain Population Covariance Structure for N=120 at .05 Level of Significance

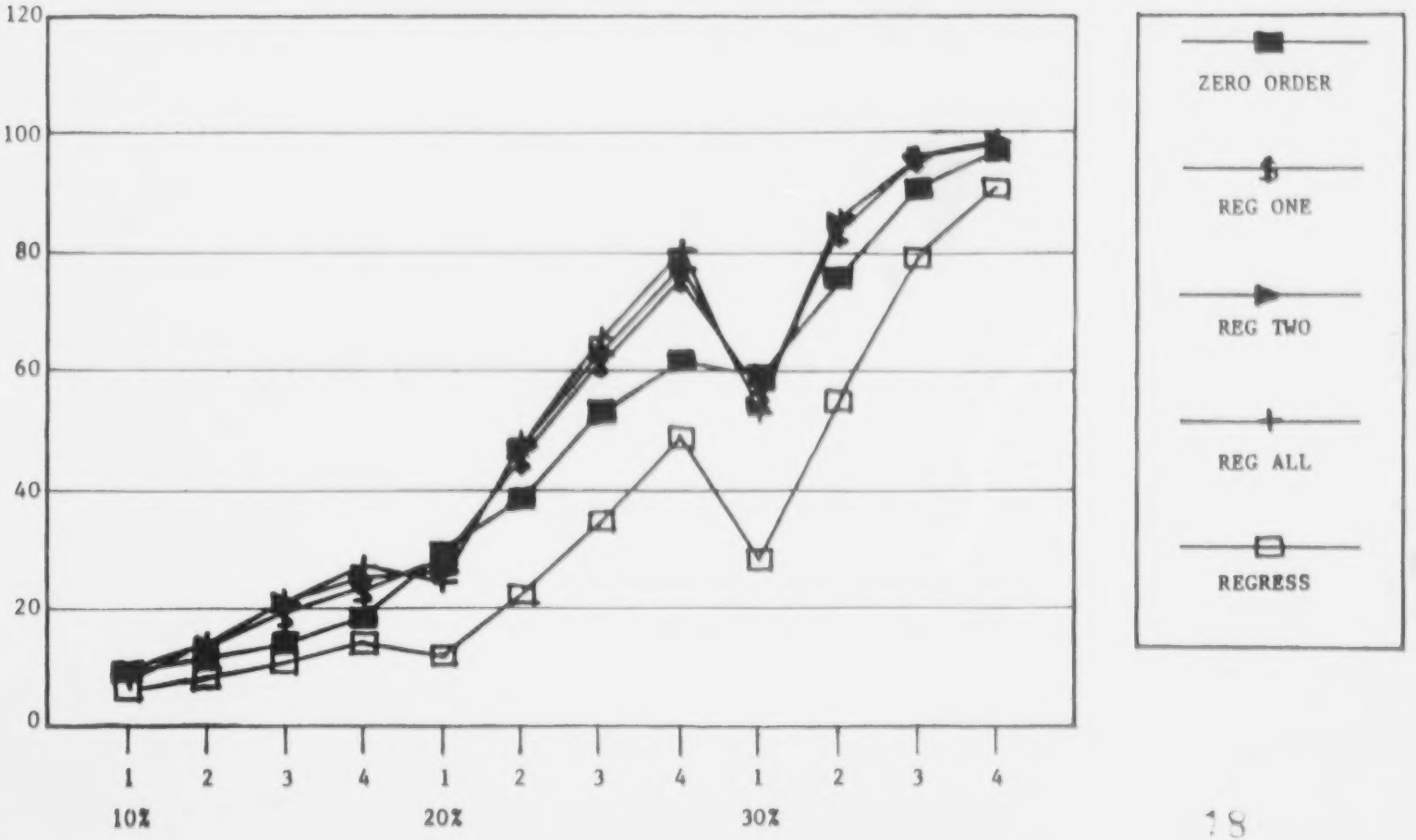

Figure 4

The Percent of Data Matrices that did not Retain Population Covariance Structure for N=30 at .10 Level of Significance

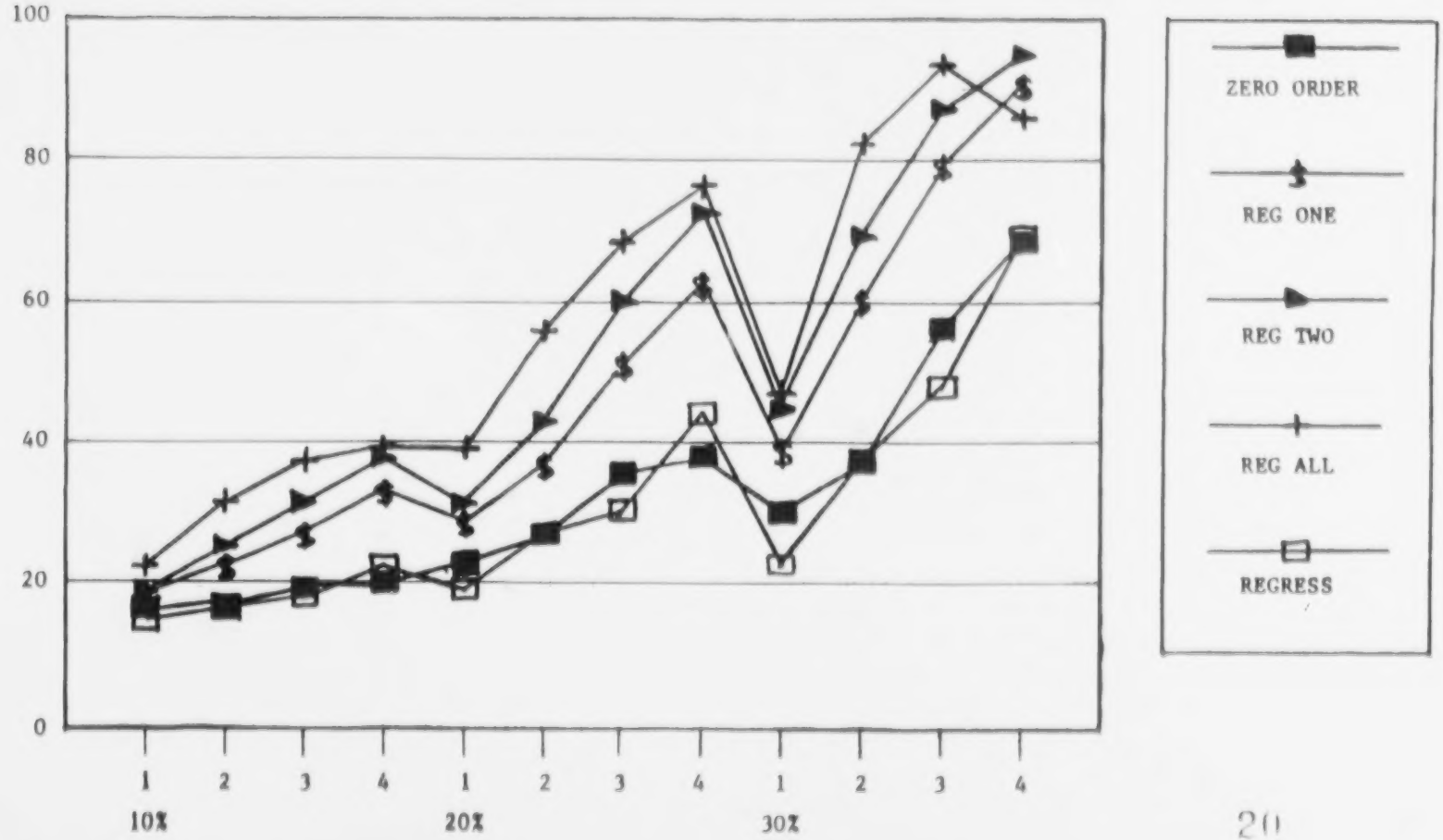

Figure 5

The Percent of Data Matrices that did not Retain Population Covariance Structure for N=60 at .10 Level of Significance

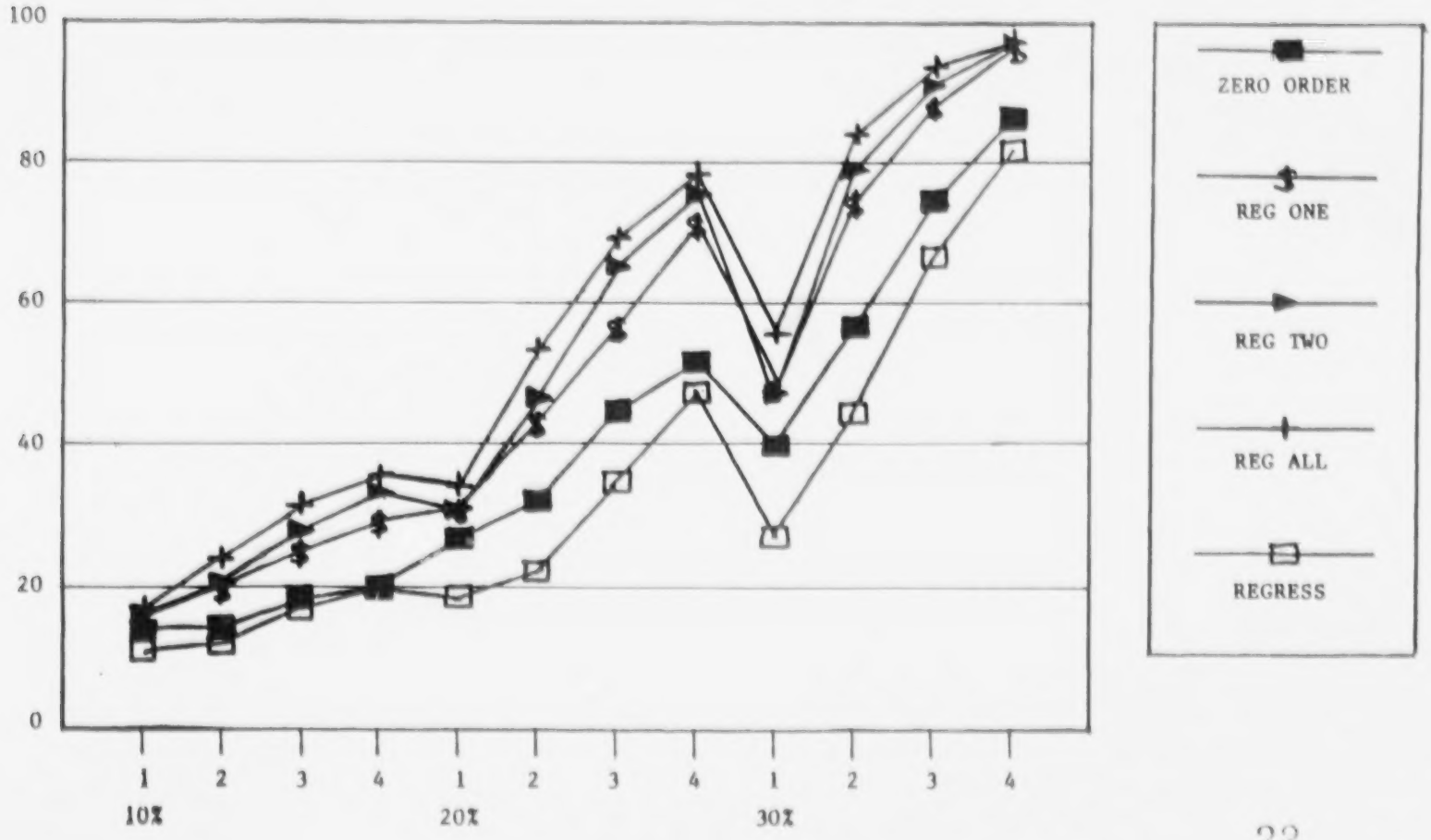

Figure 6

The Percent of Data Matrices that did not Retain Population Covariance Structure for N=120 at .10 Level of Significance